(2)

AD-A179 997

NAMRL- 1319

DEVELOPMENT OF A COMPUTER-BASED NAVAL AVIATION SELECTION TEST BATTERY

D. L. Damos and G. D. Gibb

August 1986

NAVAL AEROSPACE MEDICAL RESEARCH LABORATORY
PENSACOLA, FLORIDA

87 5 4 059

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

AD-A179 797

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS N/A
2a. SECURITY CLASSIFICATION AUTHORITY N/A	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE N/A	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) NAMRL-1319	5. MONITORING ORGANIZATION REPORT NUMBER(S) N/A
6a. NAME OF PERFORMING ORGANIZATION Naval Aerospace Medical Research Laboratory; 6b. OFFICE SYMBOL (If applicable) Code 03	7a. NAME OF MONITORING ORGANIZATION Naval Medical Research and Development Command
6c. ADDRESS (City, State, and ZIP Code) Naval Air Station, Pensacola, FL 32508-5700	7b. ADDRESS (City, State, and ZIP Code) NMC, NCR Bethesda, MD 20814-5044
8a. NAME OF FUNDING/SPONSORING ORGANIZATION; 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c. ADDRESS (City, State, and ZIP Code)	10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.
63706N	M00960.01	1051	DN477559

11. TITLE (Include Security Classification)

(U) DEVELOPMENT OF A COMPUTER-BASED NAVAL AVIATION SELECTION TEST BATTERY

12. PERSONAL AUTHOR(S)
D. L. Damos* and G. D. Gibb

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Interim	FROM 10/82 TO 8/86	86-8	24

16. SUPPLEMENTARY NOTATION
* Current address for D. L. Damos is Department of Human Factors, ISSM, University of Southern California, Los Angeles, California 90089-0021.

17. COSATI CODES		
FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)

Selection, performance measurement, computer-based testing

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Since World War II, efforts to improve tests used to select aircrew have continued. Due to the escalating costs of training aircrew, particularly pilots, improvements in the predictive validity of aircrew selection batteries have become increasingly important.

At present, the general consensus of the selection community is that existing paper-and-pencil tests fail to adequately measure four major areas of individual differences that could increase the predictive validity of aircrew selection batteries: psychomotor skills, information processing abilities, higher-order cognitive processes, and personality. This report describes a new aircrew selection battery recently developed at the Naval Aerospace Medical Research Laboratory that is intended to measure individual differences in these areas.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL J. O. Houghton, CAPT MC USN	22b. TELEPHONE (Include Area Code) (904) 452-3286; 22c. OFFICE SYMBOL Code 00

DD FORM 1473, 84 MAR — 83 APR edition may be used until exhausted. All other editions are obsolete

SECURITY CLASSIFICATION OF THIS PAGE

★U.S. Government Printing Office: 1985—507-047

UNCLASSIFIED

20. Abstract (Continued)

Two sets of data are presented for each task in the experimental battery. One set was obtained from operational aircrew; the other, from aviation officer candidates. Data are also provided indicating the differential stabilities, reliabilities, and interrelations among the dependent measures.

We recommend that the newly developed battery, after refinement, be administered to 50 aviation candidates and their performance through primary flight training monitored. Selection battery measures can then be compared to criterion measures in the flight training environment to assess the predictive validity of the various selection battery tests.

SUMMARY PAGE

THE PROBLEM

Since World War II, there has been a continuous effort to improve the tests used to select aircrew. Because of the escalating costs of training aircrew, particularly pilots, it has become increasingly important to improve the predictive validity of aircrew selection batteries. At present, the general consensus of the selection community is that existing paper-and-pencil tests fail to measure adequately four major areas of individual differences that could increase the predictive validity of aircrew selection batteries: psychomotor skills, information processing abilities, higher-order cognitive processes, and personality. This report describes a new aircrew selection battery recently developed at the Naval Aerospace Medical Research Laboratory that is intended to measure individual differences in these areas.

FINDINGS

Two sets of data are presented for each task in the experimental battery. One set was obtained from operational aircrew members; the other, from aviation officer candidates. Data are also provided indicating the differential stabilities, reliabilities, and correlations among the dependent measures. Upon minor revision of some component tests, the battery should undergo validation.

RECOMMENDATIONS

We recommend that the newly developed battery, after further refinement, be administered to 500 aviation candidates and their performance through primary flight training monitored. Selection battery measures can then be compared to criterion measures in the flight training environment to assess the predictive validity of the various selection battery tests.

Acknowledgments

The authors acknowledge the technical assistance of Peter Collyer and Kathy Vogel, who developed the computer programs for all tests. The authors also appreciate the diligent efforts of Al Thomas and Aura Walking Eagle during the field test administration.

Volunteer subjects were recruited, evaluated, and employed in accordance with the procedures specified in Department of Defense Directive 3216.2 and Secretary of the Navy Instruction 3900.39 series. These instructions are based upon voluntary informed consent and meet or exceed the provisions of prevailing national and international guidelines.

INTRODUCTION

Since World War II, there has been a continuing effort to improve the tests used to select aircrew (13). Aptitude tests and biographical inventories have been updated periodically, and new tests have occasionally been added. Despite approximately 40 years of effort, the pilot composite of the United States Naval Aviation Selection Battery, however, has an uncorrected predictive validity of approximately 0.15 to 0.25 to a pass/fail criterion for undergraduate pilot training. The Air Force Officer Qualifying Test has predictive validities that are typically in the same range (9).

Because of the escalating costs of training aircrew, it has become increasingly important to improve the predictive validity of aircrew selection batteries. At present, the general consensus of the selection community is that the existing paper-and-pencil tests fail to test adequately four major areas of individual differences that could increase the predictive validity of aircrew selection batteries: psychomotor skills, information processing abilities, higher-order cognitive processes, and personality.

The lack of psychomotor tests in the existing aircrew batteries is an historical anomaly; during World War II, both the Navy and the Army Air Corps aircrew selection batteries included extensive apparatus tests to evaluate psychomotor skills. In the early 1950's, apparatus tests were eliminated from both batteries because of problems with calibration and reliability. Sub-sequently, it was assumed that any psychomotor tests would encounter similar problems. More recently, advances in microprocessors have eliminated calibration and reliability problems and have made large-scale testing feasible. Recent studies (9, 12) of two computer-generated psychomotor tests demonstrated that scores from the two psychomotor tests made unique contributions to prediction beyond that contributed by the existing paper-and-pencil aircrew tests. Because of these encouraging results, these two tests will soon be added to the Air Force aircrew selection battery.

To date, none of the aircrew selection batteries has been constructed to test basic information processing abilities, such as reaction time and memory retrieval time. Only very limited attempts to test some higher-order cognitive processes, such as the rotation of figures in two-dimensional space have been made. Generally, basic information processing abilities have not been tested because these require measuring reaction times to millisecond accuracy. Of course, such measurement was not feasible until inexpensive microprocessors became widely available. Additionally, selection specialists were not interested in testing basic information processing abilities until some evidence was available that these processes were related to more complex behavior, such as reading speed (10) or verbal IQ scores (8). Because aircrew tasks have changed so radically since World War II, tests of basic information processing abilities, which are not related to general intelligence, could possibly improve the predictive validity of the current batteries.

Few higher-order cognitive processes have been examined in pilot selection batteries for many of the same reasons that basic information processing abilities have not been assessed: measurement requires millisecond accuracy, and there was little evidence until recently that these processes were related to more complex behavior. Furthermore, some of these processes are much more difficult to assess than basic processes because of large individual differences in performance

strategies and problems in data analysis. Because of the obvious importance of spatial manipulations and timesharing in many flight tasks, some attempt has been made to include tests of these processes in some aircrew selection batteries. Currently, tests of spatial processing are included in both the Navy and Air Force batteries; nevertheless, because both batteries use only paper-and-pencil tests, all of these tests consider only response accuracy. Egan (5) and Carter and Woldstad (3) demonstrated that the reaction times and accuracy scores of spatial tasks measure different aspects of spatial processing. The current batteries, therefore, assess only a few spatial processes at best. Neither of the current batteries contains any tests of timesharing ability although such tests were used in the Army Air Corp battery in World War II and are currently used by SAS Airlines (16).

Both the United States Navy and the Air Force have investigated a variety of personality tests for use in aircrew selection batteries (7). Because of extensive subject bias, however, personality tests have had very little impact on the selection of aircrew members. Both services are continuing to investigate personality tests, emphasizing measures of decisiveness, compulsivity, and risk taking.

This report describes a new aircrew selection battery recently developed at the Naval Aerospace Medical Research Laboratory. This battery was designed to assess basic information processing abilities, higher-order processes, psychomotor skills, and personality traits that might predict success in aviation training. The major constraint in the construction of the battery was the total testing time, which was limited to approximately 4 hours. With this restriction, we focused on a small number of basic and higher-order processes, personality traits, and psychomotor skills. Simultaneously with our effort, the Air Force Human Resources Laboratory was testing the Basic Aptitude Battery (BAT), a computer-based aircrew selection battery. Consequently, the battery described in this report was designed to complement the BAT to avoid duplication of effort.

This report presents two sets of data for each task in the battery. One set was obtained from operational aircrew; the other, from aviation officer candidates. Of interest were skill differences between experienced and entry level personnel. Analyses comparing the performance of the two groups are provided accordingly.

METHOD

Subjects. Forty-one pilots and flight officers, who were attached to the United States Navy Fighter Squadron 43 and United States Marine Corps Squadron 451, and 60 aviation candidates awaiting flight training participated in the study. Subjects were informed that the investigation involved performing tasks in problem solving, and perceptual and motor skills. They were also told that the results would not be entered into their permanent service records. The 27 pilots and 14 flight officers ranged in age from 27 to 40 (M = 30.11 years, SD = 3.03 years) and had an average of 1498.10 flight hours (SD = 769.75). Four of the pilots and flight officers were left-handed. The aviation training candidates ranged in age from 21 to 30 (M = 22.98 years, SD = 1.87 years) and had an average of 8.75 flight hours (SD = 33.90). Four of the aviation candidates were left-handed.

Apparatus. All testing was conducted on Apple IIe microcomputers with Amdek Color I Plus monitors (CRTs). Subjects used an Apple IIe numeric keypad placed under their right hand to respond to discrete stimuli. All responses were recorded to millisecond accuracy. A Measurement Systems Incorporated 542 control stick was used for cursor control during the tracking tasks. The control stick was mounted on the experimental chair and placed between the subject's legs.

Tasks. A 13-task battery measuring spatial and information processing abilities, psychomotor skills, personality traits, and dual-task performance was administered to all subjects. The first 10 tasks were administered in the order given below. These 10 tasks were followed by 3 other tasks--the psychomotor device task, the dichotic listening task, and a combination consisting of the psychomotor device and the dichotic listening tasks. The results of these last three tasks are not reported here. Additionally, the results of 1 of the first 10 tasks, the Dot Estimation Task, are described in Gibb and Lambirth (6). This report describes nine tasks in detail.

For all of the tasks described below except the single- and dual-task tracking and the time estimation tasks, the subjects were asked to respond as quickly and accurately as possible. The number of correct and incorrect responses and their associated reaction times were recorded. Because of problems with the capacity of the online data storage system, however, only summary measures were obtained for two of the tasks, maze tracing and the Baddeley Test of Grammatical Reasoning. The subjects always indicated a correct or "same" response by pressing the key under their right index finger. An incorrect or "different" response was indicated by pressing the key under their right second (middle) finger. At worst, the system required 1180 ms to erase one stimulus and present another after the subject made a response. The subjects received no performance feedback on any of the tasks, including the tracking tasks and the time estimation task. All of the tasks except single- and dual-task tracking were unpaced.

Four tasks--rotated letters, maze tracing task, grid, and manikin--were assumed to assess spatial processes. The manikin task and the rotated letters task have been found (1, 3) to be positively correlated with standardized paper-and-pencil tests of spatial abilities. The grid task is a variation of a task developed by Phillips (15) that was instrumental in proving the existence of the spatial short-term memory system. Despite the existence of supporting evidence for the spatial nature of these tasks, all tasks were pretested with a verbal suppression task to establish that spatial processes were required to perform each one of them.

Rotated letters. Two letters were presented simultaneously on the screen. The standard letter was presented in the upper right corner of the display; the comparison letter, in the bottom left corner. The comparison stimulus was rotated 0, 45, 90, 135, 180, 225, 270, or 315° relative to the standard stimulus and was identical either to the standard letter or its mirror image. If the comparison letter was a mirror image of the standard letter, the subject responded "different." If the comparison letter was simply a rotated form of the standard, the subject responded "same."

The letter "F" was used as a practice letter for this task. The subject received one presentation of a standard F and a mirror-image F at each of the eight possible rotations. The subject then received four presentations of the standard form and the mirror-image form of the letters "G" and "R" at each of the eight possible rotations. The order of stimuli for each letter was randomized with

regard to rotation and image. The task contained no rest periods and typically required 25 min. The stimuli were 2.5 cm by 2 cm and the centers of the letters were separated by 9 cm.

Dot estimation. This was a 6-min task containing a maximum of 50 trials. The CRT was divided into two fields, each 6.25 cm by 8 cm. On the first trial, one field contained one dot; the other, two dots. The subject pressed one of two keys on the keypad indicating which field had the greater number of dots. On half of the trials, the left field had the additional dot; on the other half, the right field had the additional dot. On each subsequent trial, the number of dots in each of the fields was increased by one. However, the location of the field containing the additional dot varied pseudorandomly from trial to trial. The dots were arranged in an arbitrary pattern. The number of trials attempted as well as the reaction times for correct and incorrect responses were used as dependent variables. This task is described in more detail in Gibb and Lambirth (6).

Maze tracing. In this task, 24 unique, 6.25 cm by 5.5 cm complex mazes were displayed sequentially to the subject. Each of the mazes had an entrance and an exit, but 12 of the mazes had no path from the entrance to the exit. In other words, to reach the exit, the subject had to cross at least one wall of the maze. If subjects decided the maze could be transversed normally, they pressed the key under their index finger. If they thought they had to cross at least one wall to reach the exit, they pressed the key under their second finger. Subjects were instructed not to trace through the maze manually. The task typically required 15 min.

Grid. For this task, 5 by 5 matrix grids measuring 6.5 cm by 5.5 cm were presented sequentially to the subject. Each matrix had five illuminated cells that were selected at random. The subject determined if the current matrix was identical to the preceding matrix rotated 90° to the right or left. For every trial, approximately 50% of the correct responses were "same" and 50% were "different." The same matrix could be shown sequentially (in its rotated form) a maximum of four times. The response to the first matrix pattern of any trial was always "same." This task consisted of 19 1-min trials. Trials were separated by a 30-s rest break.

Baddeley grammatical reasoning. In this task, each presentation consisted of a statement that described an order of the letters "A" and "B" followed by two letters. For example, "B follows A . . . AB." The subject determined if the sentence correctly described the order of the two letters. The sentences describing the relation between the letters involved five different grammatical transformations: 1) the use of the active versus the passive voice, 2) the veracity of the sentence, 3) the use of the verb "precedes" versus "follows," 4) affirmative versus negative phrasing, and 5) the letter mentioned first. Thirty two sentences describing the relation between the following pair of letters were possible. Each subject received one 5-min trial. The letters were 0.5 cm high.

Time estimation. For this task, the subject was required to estimate 10-s intervals as accurately as possible. The subject began estimating the first 10-s interval as soon as he saw the word "Begin" on the CRT. When he thought 10 s had elapsed, he pressed the keyboard space bar and immediately began estimating the next 10-s interval. He continued this procedure until he had estimated six 10-s intervals. After he had estimated the sixth interval, the screen was erased and a 10-s rest began. After the rest ended, a tone sounded, 2 s later the word "Begin" again appeared on the screen, and the subject began estimating the next set of six

10-s intervals. Altogether, the subject estimated three sets of intervals. The average estimated interval was the dependent measure.

Manikin. Each presentation in this task consisted of a 5 cm by 3 cm drawing of a sailor holding a square in one hand and a triangle in the other. The sailor was depicted either right side up or upside down, facing towards or away from the subject, and holding the square in either his left or right hand. Eight variations of the drawing were presented. If the sailor held the square in his left hand, the subject pressed a key under his right index finger. If the sailor held the square in his right hand, the subject pressed the key under his right second finger. Each subject completed eight 1-min trials. Each trial was separated by a 20-s rest break.

One-dimensional compensatory tracking. The subject was required to keep a 0.6-cm square centered in a 9.75 cm by 1.25 cm rectangle by making appropriate left-right movements of a control stick. The cursor was driven by a forcing function consisting of equal amplitude broadband noise. The transfer function was $Y = (.99)1/S + (.01)1/S^2$. This task was controlled by the subject's left hand. The subject received five 2-min trials. Each trial was separated by a 30-s rest. The dependent measure was RMS error. With the control stick displaced as far as possible to one side throughout the trial, the average RMS error was 125. With no control inputs, the average RMS error score was 78.

Absolute difference. In this task, randomly selected digits between one and nine were presented sequentially to the subject. The subject determined the absolute difference between the digit displayed on the CRT and the immediately preceding digit and pressed the corresponding key on the keypad. As soon as the response was entered, the digit was erased and a new one presented. All digits were presented with approximately the same frequency and a digit was never allowed to repeat. The digits were 1.25 cm by 0.5 cm and were centered inside a 1 cm by 1.25 cm rectangle. To begin the task, the subject hit any key after the first digit was displayed. Only responses of "1," "2," "3," or "4" were possible. Each subject received 10 2-min trials. Trials were separated by a 15-s rest.

Tracking-absolute difference combination. For this combination, the subject performed the tracking task and the absolute difference task concurrently. The stimuli for the absolute difference task were centered above the tracking task and touched the top of the tracking task. The subjects controlled the tracking task with their left hand and the absolute difference task with their right hand. The subject was told that the two tasks were equally important. The subject received five 2-min trials on this combination, and trials were separated by a 30 to 60-s rest. Rest intervals were variable due to data storage processing requirements. The same dependent measures were calculated for each task as under single-task conditions.

Procedure. All instructions were presented to the subjects on the CRT. The test administrators did not intervene except when problems occurred with the grid task. If a subject failed to achieve a 70% accuracy rate during the practice trial, the grid task stopped and instructions appeared to call the test administrator. Test administrators re-explained the task, and the subject performed the practice trial again. The subject repeated the practice trial until he obtained an accuracy score of at least 70%. The total administration time of the battery typically ranged from 4 to 4.5 h. Each subject saw exactly the same order of stimuli for each task. Each task was followed by a 3- to 4-min rest. During the rest break following the grid task, the subject's timepiece was removed

and returned at the conclusion of all testing. Aviation candidates were tested in a standard air-conditioned laboratory; pilots and flight officers were tested in an air-conditioned, mobile field laboratory to accommodate operational flight schedule considerations. Subjects sat at a comfortable distance from the CRT, but because of the hardware configuration, no subject could sit closer than 68 cm to the CRT.

RESULTS

Because of data storage limitations, not all of the data collected in this study could be analyzed. Consequently, pretest data for each of the tasks were examined, and the point at which 90% of the subjects reached asymptotic performance was determined. Asymptotic performance was determined for all tasks except the dot estimation, rotated letters, maze tracing task, Baddeley grammatical reasoning, and time estimation, which had single measures. For the remaining tasks, asymptotic performance was established at the beginning of the trial in which either correct reaction times or RMS error did not vary more than 10% over three consecutive trials. Except where indicated, only the post-asymptotic data from this study were analyzed. All reaction times were measured from the presentation of the stimulus to the response. The correct reaction times and percentage errors of the grid, manikin, and single-task absolute difference tasks were examined for a speed-accuracy trade-off. The correlation between the correct reaction times and the percentage error was low (.01 to .42) in all cases. Consequently, the dependent measures were always analyzed separately using univariate statistics.

The most critical statistical analyses performed on these data were conducted on the intertrial correlation matrices of the percentage correct and correct reaction times to determine both differential stability and task definition. Differential stability occurs when the group mean on a given task is not changing or is changing only in a slow linear fashion from trial to trial, the variance is constant (within some level of experimental error) from trial to trial, and the rank order of subjects is constant across trials. In a selection battery, predictions based on consistent individual differences are necessary, otherwise predictions are based on rank orders of subjects that de facto vary randomly from trial to trial. Currently, the technique used to ensure that the between-subject performance differences observed on any given trial represent a true difference is to determine if the task has become differentially stable by that trial.

Presently, two major techniques are used to estimate differential stability (2), a two-way analysis of variance and the Lawley's Test for Equality of Correlations. The analysis of variance technique compares early performance with late performance on a given task. This technique could not be used because, as noted above, performance scores from the initial trials were discarded. The usefulness of the Lawley's Test is limited in that data from a minimum of three trials must be obtained to determine stability. Consequently, estimates of differential stability could not be obtained for the maze tracing task, Baddeley, dot estimation, and time estimation tasks because only summary data were available for these tasks. Differential stability is indicated by a nonsignificant result from the Lawley's Test.

The intercorrelation matrix of stabilized trials also was examined to determine the task definition. The average intercorrelation of stabilized trials must be at least .70 for good task definition; average correlations less than this indicate that the test is too unreliable (contains excessive error variance) to be used for prediction purposes.

Before Lawley's Test was calculated on any intertrial correlation matrix, the size of the correlations in the matrix was examined. Matrices with low correlations (less than .70) were not analyzed because task definition was too poor to warrant using that dependent measure of the task as a predictor.

Dot estimation. The means and standard deviations for all dependent measures are given in Table 1. No significant differences were found between aviators and aviation candidates on the mean number of trials attempted ($F_{1,96}$ = .60), mean correct reaction time ($F_{1,96}$ = .62), or mean incorrect reaction time ($F_{1,96}$ = .22).

Table 1

Dot Estimation Task (N = 98)

	Number of Trials Attempted		Mean Correct RT (s)		Mean Incorrect RT (s)	
	M	SD	M	SD	M	SD
Aviators (n = 40)	30.75	8.38	10.36	3.14	12.25	10.45
Candidates (n = 58)	32.21	9.69	9.80	3.59	11.29	9.52

Maze tracing. Means and standard deviations for the number of correct and incorrect responses and their associated reaction times are given in Table 2. No significant differences were found between groups on the mean number of correct ($F_{1,96}$ = .06) responses, or the mean correct ($F_{1,96}$ = 1.53) or mean incorrect ($F_{1,96}$ = .37) reaction times.

Table 2

Maze Tracing Task (N = 97)

	Mean Number Correct		Mean Number Error		Mean Correct RT (s)		Mean Incorrect RT (s)	
	M	SD	M	SD	M	SD	M	SD
Aviators (n = 40)	23.00	1.09	1.00	1.09	8.22	1.76	4.89	4.22
Candidates (n = 57)	23.07	1.53	.92	1.53	7.73	2.06	4.24	5.68

Baddeley grammatical reasoning. Means and standard deviations for the dependent measures are given in Table 3. No significant differences were found between groups on the mean number of correct responses ($F_{1,96}$ = 1.04) or the mean correct ($F_{1,96}$ = .85) or mean incorrect ($F_{1,96}$ = 3.32) reaction times.

Table 3

Baddeley Grammatical Reasoning Task (N = 98)

	Mean Number Correct		Mean Number Error		Mean Correct RT (s)		Mean Incorrect RT (s)	
	M	SD	M	SD	M	SD	M	SD
Aviators (n = 40)	59.48	17.07	4.88	4.33	4.43	3.72	4.81	4.01
Candidates (n = 58)	63.03	16.87	6.12	7.75	3.84	2.45	3.52	2.95

Time estimation. The mean time estimations for aviators and aviation candidates were M = 10.90, SD = 2.09 (n = 40), and M = 11.27, SD = 2.42 (n = 51), respectively. This was not a significant difference ($F_{1,96}$ = .60).

One-dimensional compensatory tracking. The average RMS error for the final three trials of the task was used as the dependent measure. Mean RMS error was M = 20.83, SD = 6.55 and M = 30.93, SD = 12.84 for aviators (n = 40) and aviation candidates (n = 58), respectively. A one-way analysis of variance performed on the average RMS error yielded an $F_{1,96}$ = 20.94, which is significant beyond .0001.

The average intercorrelation of the last three trials was .91. Lawley's Test was significant (X^2 (2) = 7.42, p < .05), indicating that the task did not obtain stability during the testing period.

Manikin. Means and standard deviations for the number correct and incorrect and their associated reaction times averaged over the final three trials are given in Table 4. No significant differences were found between groups on the number of correct responses ($F_{1,96}$ = .06) or the reaction time for incorrect responses ($F_{1,96}$ = .24). Aviation candidates had faster reaction times for correct responses ($F_{1,96}$ = 6.71, p < .01) and more incorrect responses ($F_{1,96}$ = 3.99, p < .05).

Table 4

Manikin Task (N = 91)

	Number Correct		Number Error		Mean Correct RT (s)		Mean Incorrect RT (s)	
	M	SD	M	SD	M	SD	M	SD
Aviators (n = 40)	25.61	5.67	.96[a]	.88	1.52[b]	.49	.93	.76
Candidates (n = 51)	25.29	6.77	2.95[a]	6.27	1.30[b]	.35	.86	.55

[a] $p < .05$
[b] $p < .01$

The average intercorrelation for the correct reaction times of the last three trials was high (.82). The Lawley's Test conducted on these scores was not significant (X^2 (2) = 1.11, $p > .05$), thus, the correct reaction times were stable on the last three trials. In contrast, the average intercorrelation for the percentage correct was larger (.93), but these scores were not differentially stable (X^2 (2) = 12.82, $p < .01$).

Absolute difference. Means and standard deviations for the number correct and incorrect and their associated reaction times averaged over the last five trials are given in Table 5 for both aviators and aviation candidates. No significant differences were found between aviators and candidates on the number of correct responses ($F_{1,96}$ = 1.97) and the associated reaction time ($F_{1,96}$ = 1.21), or the reaction time for incorrect responses ($F_{1,96}$ = .80). In contrast, candidates were found to make significantly more incorrect responses ($F_{1,96}$ = 3.79, $p < .05$) than aviators.

The intercorrelation matrix of the correct reaction times showed considerable evidence of change with practice even though the average correlation was relatively high (.80). Consequently, Lawley's Test was calculated only for the last three trials. The test was significant (X^2 (2) = 21.73, $p < .01$), indicating that correct reaction times never obtained stability. The average intercorrelation for the percentage correct was too low to warrant consideration (.52).

Table 5

Absolute Difference Task (N = 98)

	Number Correct		Number Error		Mean Correct RT(s)		Mean Incorrect RT (s)	
	M	SD	M	SD	M	SD	M	SD
Aviators (n = 40)	68.06	13.64	3.29[a]	1.96	2.08	.34	2.30	.83
Candidates (n = 58)	72.39	15.90	4.96[a]	5.17	2.00	.36	2.16	.68

[a] $p < .05$

Grid. The number of correct and incorrect responses and their associated reaction times averaged over the last five trials were used as the dependent measures. Group means and standard deviations for these measures are given in Table 6. No significant differences were found between groups on the number of correct responses ($F_{1,96}$ = 2.08) or incorrect responses ($F_{1,96}$ = 0.00), or their associated reaction times ($F_{1,96}$ = 1.73) and ($F_{1,96}$ = .02), respectively.

Table 6

Grid Task (N = 98)

	Number Correct		Number Error		Mean Correct RT (s)		Mean Incorrect RT (s)	
	M	SD	M	SD	M	SD	M	SD
Aviators (n = 40)	16.41	4.91	3.00	1.29	2.19	1.66	2.38	1.34
Candidates (n = 58)	17.87	4.92	2.99	1.50	1.82	1.11	2.34	1.44

The average intercorrelation for the last five trials was .84 for the correct reaction time and .20 for the percentage correct. Since the average correlation was low for the percentage correct, no test for differential stability was conducted. The Lawley's Test performed on the correct reaction times indicated that this measure was stable for the last four trials (X^2 (5) = 5.53, $p > .05$).

Rotated letters. To obtain rates of rotation for the standard and mirror images, we averaged the data from the test stimuli "G" and "R" for the 45° and the 315° rotations, from the 90° and 270° rotations, and the 135° and 225° rotations. The slopes and intercepts for standard and mirror images were calculated on a subject-by-subject basis using the scores obtained from the averaged rotations and the data from the 180° rotation. The average amount of variance accounted for by the regression equations was 84% for standard images and 64% for mirror images. The data from the 0° rotation were not used because data obtained from this degree of rotation frequently do not appear to be functionally related to data obtained at other degrees of rotation. Regression equations that included the 0° rotation decreased the average amount of variance accounted for to 80% for standard images and 13% for mirror images. Table 7 shows the mean slopes and intercepts for aviators and aviation candidates by image. For aviators, the correlation between slopes and intercepts for the standard images was -.62. The corresponding correlation for mirror images was - .54. For aviation candidates, the correlation was -.63 for standard images and -.39 for mirror images. Because only one estimate was obtained for the slope and intercept for the standard and mirror images, no estimates of differential stability could be obtained for these measures.

Table 7

Rotated Letters Task (N = 96)

	Slope (Standard) (s)		Slope (Mirror) (s)		Intercept (Standard) (s)		Intercept (Mirror) (s)	
	M	SD	M	SD	M	SD	M	SD
Aviators (n = 40)	.245	.171	.202	.230	.641	.271	1.284	.621
Candidates (n = 56)	.278	.166	.214	.142	.603	.298	1.301	.551

Tracking-absolute difference combination. The final three trials of the task were analyzed for each group, and the means and standard deviations of the average values are given in Table 8. Mean RMS error was 27.00, SD = 8.29 for aviators, and 37.02, SD = 15.11 for aviation candidates. Candidates performed at a significantly lower level than aviators on the tracking task ($F_{1,96}$ = 14.60, p < .0002). No significant differences were found between the groups on any of the absolute difference task measures; number of correct responses ($F_{1,96}$ = .45) or incorrect responses ($F_{1,96}$ = .95), mean correct ($F_{1,96}$ = .12) or incorrect reaction time ($F_{1,96}$ = 1.45).

Although the average intercorrelation of the RMS scores was relatively high (.85), the scores did not obtain differential stability (X^2 (2) = 29.76, p < .05). The correct reaction times had a smaller average intercorrelation (.78) but were stable (X^2 (2) = 0.21, p > .05). The average percentage correct was too low (.56) to warrant examination for stability. Correlations between the tracking task and

the absolute difference task were calculated for each of the last three trials by group to determine if different strategies were used to perform the combination. No evidence of different strategies was found.

Table 8

Tracking-Absolute Difference Combination (N = 98)

	n	Mean Number Correct		Mean Number Error		Mean Correct RT (s)		Mean Incorrect RT (s)		RMS	
		M	SD	M	SD	M	SD	M	SD	M	SD
Aviators (n = 40)	40	69.27	16.15	7.43	15.12	2.01	.39	2.33	.75	27.00[a]	8.29
Candidates (n = 58)	58	66.94	17.30	10.49	15.33	1.98	.39	2.55	.93	37.04[a]	15.11

[a] $p = < .0002$

Spatial task interrelations. Pearson correlation coefficients between the maze tracing, manikin, grid, and rotated letters tasks were calculated on the percentage correct and reaction times of correct responses to determine if the tasks were significantly related. Table 9 presents the intercorrelation matrix for correct reaction times. The data indicate a strong relation between the maze tracing task with both manikin Trial 6 ($p < .01$) and Trial 7 ($p < .05$) and the rotated letters mirror image intercept ($p < .05$) tasks. The rotated letters mirror image intercept also has a significant correlation with manikin Trials 6 and 7 ($p < .03$) and manikin Trial 8 ($p < .05$). The rotated letters standard image intercept was also significantly correlated with manikin Trial 8 ($p < .05$). No other between-task coefficients reached significance for the reaction time measure.

Table 9

Intercorrelation Matrix of Correct Reaction Times Between Maze Tracing, Grid, Manikin, and Rotated Figures Tasks ($\underline{N}$ = 101)

	Maze Tracing	Grid Trial 16	Grid Trial 17	Grid Trial 18	Grid Trial 19	Manikin Trial 6	Manikin Trial 7	Manikin Trial 8	Rotated Letters Standard Orientation Intercept	Rotated Letters Standard Orientation Slope	Rotated Letters Mirror Image Intercept	Rotated Letters Mirror Image Slope
Maze Tracing		.01	.02	.07	.03	.26^{b}	.20^{d}	.15	.09	.09	.21^{d}	.08
Grid Trial 16			.86^{a}	.82^{a}	.88^{a}	-.01	.02	-.02	.07	.05	.09	-.02
Grid Trial 17				.84^{a}	.88^{a}	.00	.02	-.04	.06	.00	.11	-.11
Grid Trial 18					.87^{a}	.09	.09	.05	.04	.14	.14	-.05
Grid Trial 19						-.04	-.01	-.03	.11	.01	.08	-.05
Manikin Trial 6							.83^{a}	.80^{a}	.14	.01	.26^{c}	-.12
Manikin Trial 7								.81^{a}	.16	.00	.30^{c}	-.16
Manikin Trial 8									.21^{d}	-.05	.21^{d}	-.09
Rotated Letters Standard Orientation Intercept										.63^{a}	.31^{c}	-.25^{c}
Rotated Letters Orientation Slope											.21^{d}	.37^{b}
Rotated Letters Mirror Image Intercept												-.47^{a}
Rotated Letters Mirror Image Slope												

[a] $\underline{p} < .0001$
[b] $\underline{p} < .001$
[c] $\underline{p} < .03$
[d] $\underline{p} < .05$

With respect to percentage correct, several high correlations emerged. Table 10 presents the intercorrelation matrix for these data. Three of the four grid task trials correlated with the maze tracing task at the .05 level or better while the remaining grid task measure approached significance ($p < .07$). Of 12 possible correlations between the grid and manikin tasks, 9 correlations were significant ($p < .05$) while the remaining 3 correlations approached significance ($p < .10$). Several of these were significant beyond the .01 level.

Table 10

Intercorrelation Matrix of Percent Correct Between Maze Tracing, Grid, Manikin, and Rotated Figures Tasks (N = 101)

	Maze Tracing	Grid Trial 16	Grid Trial 17	Grid Trial 18	Grid Trial 19	Manikin Trial 6	Manikin Trial 7	Manikin Trial 8	Rotated Letters Standard Orientation	Rotated Letters Mirror Image
Maze Tracing		.20[e]	.19[e]	.28[b]	.17	.16	.13	.12	.67[a]	.23[d]
Grid Trial 16			.52[a]	.46[a]	.41[a]	.15	.19[e]	.20[e]	.24[c]	-.03
Grid Trial 17				.48[a]	.34[b]	.23[d]	.23[d]	.25[c]	.26[c]	.00
Grid Trial 18					.41[a]	.20[e]	.16	.17	.29[b]	-.09
Grid Trial 19						.26[c]	.25[c]	.24[c]	.29[b]	-.08
Manikin Trial 6							.96[a]	.95[a]	.21[d]	-.05
Manikin Trial 7								.97[a]	.17	-.05
Manikin Trial 8									.17	-.04
Rotated Letters Standard Orientation										.25[c]
Rotated Letters Mirror Image										

[a] $p < .0001$
[b] $p < .005$
[c] $p < .01$
[d] $p < .03$
[e] $p < .05$

Additionally, all four grid task measures were highly correlated with the rotated letters standard image task ($p < .01$). Half of these correlations were significant beyond the .005 level. The rotated letters standard orientation image was also highly correlated with the maze tracing task ($r = .67$, $p < .0001$). The rotated letters standard image was also correlated with manikin Trial 6 ($p < .03$),

while the remaining two correlations with manikin Trials 7 and 8, approached significance ($p < .08$).

The rotated letters mirror image was only found to correlate with the maze tracing task ($p < .03$). No other significant relations were observed between tasks for the percentage correct.

DISCUSSION

As noted earlier, differential stability is one of the major criteria for choosing tasks for a selection battery. Because of storage limitations, data from three tasks could not be collected in a manner that permitted subsequent calculation of the differential stability of the task. Additionally, not enough data were available to perform differential stability calculations on the slopes and intercepts of the rotated letters task. Analyses of the remaining four tasks and the combination indicated that only the correct reaction times from the manikin and grid tasks and the dual-task absolute difference task reached differential stability. Generally, the intertrial correlations for the percentage correct were so low that the authors made no attempt to determine if the percentage correct was differentially stable. The only exception to this was the manikin task; although the intertrial correlations were high for this task, the percentage correct did not obtain differential stability.

The reason why the intertrial correlations for the percentage correct were so low for the grid and single- and dual-task absolute difference tasks is difficult to understand. As noted by Jones (11), low intertrial correlations indicate poor reliability in the classical test theory sense and poor task definition. That is, what is being measured by the percentage correct changes from trial to trial. For percentage correct, poor intertrial correlations usually occur when some type of performance ceiling is reached. In this experiment, however, the subjects did not appear to reach a performance ceiling on any task in the battery. Thus, the most common reason for the low intercorrelations does not appear to be applicable, and no explanation is offered for them at this time.

Differential stability and high task definition can often be obtained for a given task by increasing the amount of practice on that task. Thus, by providing more trials on each of the tasks that were unstable, some probability exists that the tasks would become stable and that the intertrial correlations would increase. The test battery will be refined to include five additional 2-min trials to both single-task tracking and absolute difference tasks in an attempt to obtain differential stability and high task definition for those tasks. This, of course, raises problems with the total length of future testing sessions. The total testing time could be kept approximately the same as it is now if some of the tasks measure the same processes. If this were the case, then some of the tasks measuring the same processes could be eliminated. Four of the tasks--the maze tracing, the manikin, the grid, and the rotated letters--supposedly measure spatial processes. An examination of Tables 9 and 10, however, shows only two between-task correlations above .30. Thus, these four tasks do not appear to measure the same attributes, and the identification of redundant measures of the same processes must await subsequent analyses.

Of the nine tasks in the battery, the rotated letters task is the only one that clearly needs major changes. As noted earlier, including reaction times for the 0° rotation in the regression equation decreased, rather than increased, the

amount of variance accounted for by the equation. This is a rather common finding and occurs because the mean correct reaction times for the 0^{o} rotation do not lie on the regression line defined by the other degrees of rotation (4). Comparable results occur with the choice reaction time and the Sternberg memory search tasks and indicate a problem with our current understanding of some comparison processes rather than a problem with the implementation of the task. Generally, the slope and the intercept of this task measure different processes, as they do in the Sternberg memory search task and the choice reaction time task. The slope and the intercept of the standard images and of the mirror images should theoretically correlate .00. In this study, the slope and the intercept of the standard images correlated -.63 for aviation candidates; the correlation for mirror images was -.39. Similar correlations were obtained with aviators. We conclude that this task does not appear to be functioning correctly and may not be measuring the correct processes. To correct this problem, new versions of this task must be programmed and tested.

Between-group differences were found on four of the nine tasks in the battery: absolute difference, tracking-absolute difference combination (tracking only), manikin, and one-dimensional compensatory tracking. For the one-dimensional compensatory tracking performed alone and in combination with the absolute difference task, aviation candidates were found to have substantially greater RMS scores. Indeed, as a result of their aviation experience, aviators were expected to demonstrate superior performance on the tracking tasks. Prior flight experience has been shown to contribute to better performance on tracking tasks (14).

Aviation candidates made a greater number of errors on both the manikin and absolute difference tasks than aviators while demonstrating lower correct reaction times on the manikin task. Group differences on the manikin task may be attributed to differences in the speed/accuracy trade-off strategy of the two groups.

CONCLUSIONS

Generally, correct reaction times appeared to be the most stable of the dependent measures and would provide the most useful assessment of abilities. As stated earlier, the battery will be refined to include additional practice trials for both single-task tracking and the absolute difference tasks to obtain differential stability. The probability is high that minor revisions to these two tasks will improve their usefulness. The rotated letters task, however, requires major modifications and will be removed from the battery. These revisions will be accomplished before validation of the selection test battery begins.

We anticipate that the refined version of the selection test battery will be administered to 500 incoming naval aviator and naval flight officer candidates before they begin flight training. The predictive validity of all selection battery tasks will be determined as criterion data from the training environment become available. The candidates will have been screened by current Navy selection techniques (AQT/FAR) and will have graduated from a 16-week basic training course before testing. Thus, the candidates will be a restricted sample of applicants for naval flight training. Because the predictive validity of test scores from restricted samples is always less than that obtained from unrestricted samples, the true predictive validity of this selection battery will be underestimated.

RECOMMENDATIONS

We recommend that the newly developed battery, after it is refined, be administered to 500 aviation candidates and their performance through primary flight training monitored. Selection battery measures can then be compared to criterion measures in the flight training environment to assess the predictive validity of the various selection battery tests.

REFERENCES

1. Berg, C., Hertzog, C., & Hunt, E. 1982. Age differences in the speed of mental rotation. Developmental Psychology 18:95-107.

2. Bittner, A. 1979. Statistical tests for differential stability. Proceedings of the Human Factors Society 23rd Annual Meeting. Boston, MA:541-545.

3. Carter, R., & Wolstad, J. 1985. Repeated measurements of spatial ability with the manikin test. Human Factors 27:209-220.

4. Cooper, L. A., & Shepard, R. N. 1978. Transformation on representations of objects in space. Chapter in Handbook of Perception. Vol. VIII, New York: Academic Press.

5. Egan, D. 1978. Characterizing spatial ability: Different mental processes reflected in accuracy and latency scores. NAMRL Research Report 1250. Pensacola, FL: Naval Aerospace Medical Research Laboratory.

6. Gibb, G., & Lambirth, T. 1986. Evaluation of a behavior based personality test. Presented at the 57th Annual Meeting of the Aerospace Medical Association, Nashville, TN.

7. Griffin, G., & Mosko, J. 1977. A review of naval attrition research 1950-1976: A base for the development of future research and evaluation. NAMRL Research Report 1237. Pensacola, FL: Naval Aerospace Medical Research Laboratory.

8. Hunt, E., Frost, N., & Lunneborg, C. 1973. Individual differences in cognition: A new approach to intelligence. In G. Bower (Ed.), The Psychology of Learning and Motivation. New York: Academic Press.

9. Hunter, D., & Thompson, N. 1978. Pilot selection system development. Technical Report No. TR-78-33. San Antonio, TX: Brooks Air Force Base, Personnel Research Division.

10. Jackson, M., & McClelland, R. 1979. Processing determinants of reading speed. Journal of Experimental Psychology: General 108:151-181.

11. Jones, M. B. 1979. Stabilization and task definition in a performance test battery. Proceedings of the Human Factors Society Annual Meeting. Santa Monica, CA:536-540.

12. Kantor, J., & Bordelon, V. 1985. The USAF pilot selection and classification research program. Aviation, Space, and Environmental Medicine 56:254-257.

13. North, R., & Griffin, G. 1977. Aviator selection 1919-1977. NAMRL Special Report 77-2. Pensacola, FL: Naval Aerospace Medical Research Laboratory.

14. Petho, F. C. 1983. Prior flight experience as a moderator variable in a complex psychomotor performance task. Preprints of the Scientific Program of the Aerospace Medical Association Annual Meeting, 187, 1983.

15. Phillips, W. 1974. On the distinction between sensory storage and short-term visual memory. Perception and Psychophysics 16:283-290.

16. Trankell, A. 1959. The psychologist as an instrument of prediction. Journal of Applied Psychology 43:170-175.